SF 298 MASTER COPY   KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1998	Technical - 98-01

4. TITLE AND SUBTITLE
Second Order Corrections of the Sequential Bootstrap

5. FUNDING NUMBERS
DAAH04-96-1-0082

6. AUTHOR(S)
G.J. Babu, P.K. Pathak and C.R. Rao

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Center for Multivariate Analysis
417 Thomas Bldg.
Department of Statistics
Penn State University
University Park, PA 16802

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 35518.31-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

ABSTRACT. Rao, Pathak and Koltchinskii (1997) have recently studied a sequential approach to resampling in which resampling is carried out sequentially one-by-one (with replacement each time) until the bootstrap sample contains $m \approx (1-e^{-1})n \approx .632n$ distinct observations from the original sample. They have established that the main empirical characteristics of the sequential bootstrap go through, in the sense of being within a distance of order $O(n^{-3/4})$ from those of the usual bootstrap. However, the theoretical justification of the second order correctness of the sequential bootstrap is somewhat involved. It is the main topic of this investigation. Among other things, we accomplish it by approximating our sequential scheme by a resampling scheme based on the Poisson distribution with mean $\mu = 1$ and censored at $X = 0$.

14. SUBJECT TERMS
Asymptotic expansions, Edgeworth expansions, bootstrap, expansions for conditional distrobutions, lattice distribution

15. NUMBER IF PAGES
16

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

SECOND ORDER CORRECTIONS OF THE SEQUENTIAL BOOTSTRAP

Gutti Jogesh Babu, P.K. Pathak, and C.R. Rao

Technical Report 98-01

May 1998

Center for Multivariate Analysis
417 Thomas Building
Penn State University
University Park, PA 16802

Research work of authors was supported by the Army Research Office under Grant DAAHO4-96-1-0082. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

SECOND ORDER CORRECTIONS OF THE SEQUENTIAL BOOTSTRAP

GUTTI JOGESH BABU, P. K. PATHAK, AND C. R. RAO

ABSTRACT. Rao, Pathak and Koltchinskii (1997) have recently studied a sequential approach to resampling in which resampling is carried out sequentially one-by-one (with replacement each time) until the bootstrap sample contains $m \approx (1-e^{-1})n \approx .632n$ distinct observations from the original sample. They have established that the main empirical characteristics of the sequential bootstrap go through, in the sense of being within a distance of order $O(n^{-3/4})$ from those of the usual bootstrap. However, the theoretical justification of the second order correctness of the sequential bootstrap is somewhat involved. It is the main topic of this investigation. Among other things, we accomplish it by approximating our sequential scheme by a resampling scheme based on the Poisson distribution with mean $\mu = 1$ and censored at $X = 0$.

1. INTRODUCTION

Efron (1979) introduced the bootstrap method of resampling as a ubiquitous sampling technique of estimating the variance of an estimator and a sampling distribution of a given statistic. In a fundamental paper, Bhattacharya and Ghosh (1978) have demonstrated that Edgeworth expansions for a wide class of statistics can be derived from Edgeworth expansions for multivariate sample means. This technique has been used by Singh (1981) to show, in the case of univariate sample mean, that the bootstrap is more accurate than the central limit theorem when higher order population moments exist. These ideas are further exploited by Babu and Singh (1983, 1984) to show the superiority of the bootstrap method and by Babu and Singh (1985) to obtain Edgeworth expansions for the ratio statistic and similar statistics based on

Date: December 15, 1997.

1991 *Mathematics Subject Classification.* 62G09, 62E20, 60F05.

Key words and phrases. Asymptotic expansions, Edgeworth expansions, bootstrap, expansions for conditional distributions, lattice distribution.

Research work of Gutti Jogesh Babu was supported in part by NSA grant MDA904-97-1-0023 and NSF grant DMS-9626189. Research work of C. R. Rao was supported by the Army Research Office under Grant DAAH04-96-1-0082.

samples from finite populations. The method is also used by Babu and Singh (1989) to obtain global Edgeworth expansions for functions of means of random vectors, when one of the coordinates has a lattice distribution and the remaining part of the vector has a strongly non-lattice distribution. Later Gene and Zinn (1990) showed that in a certain weak sense, the bootstrap method is valid (consistent) if and only if the central limit theorem holds. In fact the central limit theorem furnishes accuracy of approximation of order $o(1)$, while if the third population moment exists, one can expect, in many commonly encountered populations the accuracy of the bootstrap method to be of order $o(n^{-1/2})$, where n denotes the sample size. Thus while the bootstrap method has the potential of being second-order accurate; the central limit approximation is not so. This is one of the several reasons for the current interest and preference in the literature for those methods of resampling that are second-order accurate, i.e., accurate of the order $o(n^{-1/2})$.

Stemming from Efron's observation that the information content of a bootstrap sample is based on approximately $(1 - e^{-1})100\% \approx 63\%$ of the original sample, Rao, Pathak and Koltchinskii (1997) have introduced a sequential resampling method in which sampling is carried out one-by-one (with replacement) until $(m + 1)$ distinct original observations appear, where m denotes the largest integer not exceeding $(1 - e^{-1})n$. The last observation is discarded to ensure simplicity in technical details. It has been shown that the empirical characteristics of this sequential bootstrap are within a distance of order $O(n^{-3/4})$ from the usual bootstrap. The authors provide a heuristic argument in favor of their sampling scheme and establish the consistency of the sequential bootstrap; however the question of second-order correctness was not addressed.

One of the main advantages of the sequential bootstrap over the classical fixed sample size bootstrap is its performance in estimating the variance of an estimator, when the original data contains several identical values. This situation occurs when the sample is drawn from a population with a distribution that is not continuous. To be more specific, suppose $X_1, \ldots, X_n$ are i.i.d random variables satisfying $P(X_1 = 0) = .3$. Then with positive probability, about 30% of the data X_i are equal to 0.

With positive probability, several bootstrap resamples end up in n zeros, leading to a zero estimate of variance. On the other hand the sequential bootstrap, by sampling until m distinct labels X_i are selected, guarantees a resample that contains elements other than 0. Hence the sequential bootstrap scheme has an edge over the classical bootstrap, especially when dealing with categorical data.

The main object of this paper is to examine the second order correctness of the sequential bootstrap. The theoretical justification of this is somewhat more difficult because of the dependence among the bootstrap sample units. At this time a rigorous Edgeworth expansion under this kind of dependence is unavailable in the literature. A cumbersome approach based on computation of cumulants, under the (unsubstantiated) assumption that a formal Edgeworth expansion is valid, may be given along the lines of the Hall-Mammen (1994) paper. This does not lead to a complete solution as the Edgeworth expansions are not known. Instead we first approximate the sequential bootstrap by another sequential resampling scheme based on the Poisson distribution. Under the new scheme the "independence" of sample units under resampling is preserved. A rigorous justification of the Edgeworth expansion can now be given more easily. This then provides a sound theoretical framework in which the second order correctness for the sequential bootstrap can be established. In this paper we concentrate on sample means of k-variate random vectors. The Edgeworth expansions for smooth functions of multivariate sample means follow from similar expansions for multivariate means as in Bhattacharya and Ghosh (1978).

2. Sequential Resampling Schemes

Let $S = (X_1, X_2, \ldots, X_n)$ be a random sample from a distribution F, and $\theta(F)$ a parameter of interest. Let F_n denote the empirical distribution function based on S, and suppose that $\theta(F_n)$ is to be used as an estimator of $\theta(F)$. The Efron's bootstrap method approximates the sampling distribution of a standardized version of $\sqrt{n}(\theta(F_n) - \theta(F))$ by the resampling distribution of a corresponding statistic $\sqrt{n}(\theta(\hat{F}_n) - \theta(F_n))$ based on a bootstrap sample $\hat{S}_n$ in which the original F has been replaced by the empirical distribution based on the original sample S, and F_n of the

former statistic has been replaced by the empirical distribution based on a bootstrap sample $\hat{F}_n$. In Efron's bootstrap resampling scheme, $\hat{S}_n = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)$ is a random sample of size n drawn from S by simple random sampling with replacement (SRSWR). In the Rao-Pathak-Koltchinskii (1997) sequential scheme, observations are drawn from S sequentially by SRSWR until there are $(m+1) = [n(1-e^{-1})] + 2$ distinct original observations in the bootstrap sample; the last observation is discarded to ensure technical simplicity. Thus an observed bootstrap sample under the Rao-Pathak-Koltchinskii scheme admits the form:

$$\hat{S}_N = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) \tag{2.1}$$

in which $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N$ have $m \approx n(1-e^{-1})$ distinct observations from S. The random sample size N admits the following decomposition in terms of the independent random variables:

$$N = I_1 + I_2 + \ldots + I_m \tag{2.2}$$

in which $m = [n(1-e^{-1})] + 1$; $I_1 = 1$, and for each k, $2 \le k \le m$,

$$P(I_k = j) = \left(1 - \frac{k-1}{n}\right)\left(\frac{k-1}{n}\right)^{j-1}. \tag{2.3}$$

Although we have established the consistency of this sampling scheme, a rigorous proof of its second order correctness requires an Edgeworth expansion for dependent random variables; such an expansion is unavailable in the literature at the present time. An alternative approach that can be used is to slightly modify the preceding resampling scheme so that existing techniques on Edgeworth expansion, such as those of Babu and Bai (1996), Bai and Rao (1992), Babu and Singh (1989) and others, can be employed. A modification of our previous resampling scheme that allows the second-order correctness to go through easily is as follows:

Poisson Resampling Scheme:

For the selection of a bootstrap sample with a given number m of distinct units, under the Poisson Resampling Scheme (PRS), we provide a conceptual definition and a practical approach. Let us take a sample $\alpha_1, \ldots, \alpha_n$ of n independent observations

from $P(1)$, i.e., Poisson distribution with mean 1. If there are exactly m values in the sample, we accept it and take

$$\hat{S} = \{(X_1, \alpha_1), (X_2, \alpha_2), \dots, (X_n, \alpha_n)\}, \tag{2.4}$$

i.e., with the observation X_i repeated α_i times, as the bootstrap sample. If the number of nonzero values in $\alpha_1, \dots, \alpha_n$ is not exactly m, we reject the entire sample and draw another sample of size n. The bootstrap sample size N of $\hat{S}$ as in (2.1) is a random variable

$$N = \alpha_1 + \ldots + \alpha_n. \tag{2.5}$$

A practical way of implementing this resampling scheme is to first assign at random $(n - m)$ α's a value of zero and to the remaining m α's values independently chosen from the Poisson distribution with mean $\mu = 1$ and censored at $X = 0$. An outline of the equivalence of these two procedures is as follows.

Theorem 2.1. *The moment generating function $M_N(t)$ of N, the sample size of the Poisson resampling scheme, is given by*

$$M_N(t) = \left[\frac{(e^{(e^t-1)} - e^{-1})}{(1 - e^{-1}}\right]^m. \tag{2.6}$$

Proof. Let $Y_1, Y_2, \dots, Y_n$ be n Poisson variables with mean $\mu = 1$. Then it is easily seen that

$$P(N = w) = \text{const}\left(\Sigma_1 \frac{e^{-n}}{\alpha_{1!}\alpha_{2!}\ldots\alpha_{m!}}\right) \tag{2.7}$$

where the sum Σ_1 extends over all positive natural numbers $\alpha_1, \alpha_2, \dots, \alpha_m$ such that $\alpha_1 + \alpha_2 + \ldots + \alpha_m = w$. It then follows that (see Pathak (1961))

$$\begin{aligned} P(N = w) &= \text{const}\left[\frac{e^{-n}}{w!}(m^w - \binom{m}{1}(m! - 1)^w + \ldots \pm 1^w)\right] \\ &= \text{const}\frac{e^{-n}}{w!}(\Delta^m X^w|_{X=0}) \end{aligned} \tag{2.8}$$

where Δ is the difference operator with unit increment.

From (2.8) it follows that

$$P(N = w) = \frac{1}{(e-1)^m}\Delta^m \frac{X^w}{w!}|_{X=0}. \tag{2.9}$$

Consequently the moment generating function $M_N(t)$ of N is given by

$$
\begin{aligned}
M_N(t) = E(e^{tw}) &= \sum_{w\geq 0} \frac{e^{tw}}{(e-1)^m} \Delta^m \frac{X^w}{w!}|_{X=0} \\
&= \frac{1}{(e-1)^m} \sum_{w\geq 0} \Delta^m \frac{(e^t X)^w}{w!}|_{X=0} \\
&= \Delta^m \frac{e^{Xe^t}}{(e-1)^m}|_{X=0} \qquad (2.10) \\
&= \frac{1}{(e-1)^m}\{e^{me^t} - \binom{m}{1} e^{(m-1)e^t} + \binom{m}{2} e^{(m-2)e^t} \ldots\} \\
&= \frac{e^{me^t}}{(e-1)^m}\{1 - \binom{m}{1} e^{-t} + \binom{m}{2} e^{-2t} - \ldots\} \\
&= \frac{e^{me^t}}{(e-1)^m}(1 - e^{-e^t})^m \\
&= \left[\frac{(e^{e^t}-1)}{(e-1)}\right]^m \\
&= \left[\frac{(e^{(e^t-1)} - e^{-1})}{(1-e^{-1})}\right]^m \qquad (2.11)
\end{aligned}
$$

This completes the proof. □

The preceding theorem shows that the distribution of N can be viewed as that of the sum of m IID random variables with a common distribution with the moment generating function given by the formula

$$
m(t) = \frac{(e^{(e^t-1)} - e^{-1})}{(1-e^{-1})} \qquad (2.12)
$$

It is evident that $m(t)$ is the moment generating function of the Poisson distribution with $\mu = 1$ and censored at $X = 0$. Let Y denote a random variable with moment generating function $m(t)$. Then $E(Y) = 1/(1-e^{-1})$ and $V(Y) = e(e-2)/(e-1)^2$. Therefore

$$
\begin{aligned}
E(N) &= mE(Y) \\
&= n + 0(1) \qquad (2.13)
\end{aligned}
$$

and

$$\begin{aligned} V(N) &= mV(Y) \\ &= n(e-2)/(e-1) + 0(1) \end{aligned} \tag{2.14}$$

With these results, we now proceed to establish the second order correctness of the sequential bootstrap based on the Poisson distribution.

3. Second Order Correction

Let $\{a_{1,n}, \ldots, a_{n,n}\}$ be a sequence of column vectors in $\boldsymbol{R}^k$. In the application we typically use

$$a_{i,n} = (X_i - \bar{X}) \quad \text{with } \bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i. \tag{3.1}$$

Note that k denotes the dimension of X_i. Let $\{Y_j : j \geq 1\}$ be a sequence of non-negative i.i.d. random variables with a lattice distribution of span 1. Let

$$\begin{aligned} &\mu = E(Y_1), \quad \sigma^2 = V(Y_1) > 0, \\ &\gamma_3 = E(Y_1 - \mu)^3\sigma^{-3}, \quad p = P(Y_1 > 0), \quad q = 1 - p. \end{aligned}$$

Further assume that the support of Y_1 has at least two non-zero values. Note that $\gamma_3 = 1$ if Y_1 has a Poisson distribution with mean $\mu = 1$. Let

$$V_n^2 = \frac{1}{n}\sum_{j=1}^n a_{j,n}a'_{j,n}, \tag{3.2}$$

$$c_{j,n} = V_n^{-1}a_{j,n} \tag{3.3}$$

$$p_n(x) = \frac{1}{6n}\sum_{j=1}^n((c'_{j,n}x)^3 - 3(c'_{j,n}\mathbf{1})^2(c'_{j,n}x)), \tag{3.4}$$

where $\mathbf{1}$ denotes the column vector in $\boldsymbol{R}^k$ with all the entries equal to 1, and for any vector $t \in \boldsymbol{R}^k$, let

$$d_n(t) = \frac{1}{n}\sum_{j=1}^n e^{it'a_{j,n}}. \tag{3.5}$$

Define for any measurable function h on $\boldsymbol{R}^k$,

$$M_h = \sup|h(x)|(1 + \|x\|)^{-3}, \tag{3.6}$$

and for any $\delta > 0$, $x \in \boldsymbol{R}^k$,

$$w(h,\delta;x) = \sup_{\|x-z\|<\delta} |h(x) - h(z)|,$$

$$w(h,\delta) = \int_{\boldsymbol{R}^k} w(h,\delta;x)\phi_k(x)\,dx, \tag{3.7}$$

where ϕ_k denotes the density of the k-variate standard normal distribution.

Further define

$$N = \sum_{i=1}^{n} Y_i,\ \bar{Y} = N/n \tag{3.8}$$

$$U_n = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} Y_j V_n^{-1} a_{j,n} \tag{3.9}$$

$$T_n = \sum_{j=1}^{n} I_{\{Y_j > o\}}, \tag{3.10}$$

and

$$\tilde{\Psi}_n(x) = \left(1 + \frac{1}{\sqrt{n}}\gamma_3 P_n(x)\right)\phi_k(x). \tag{3.11}$$

Let $F_n^0(\cdot|m)$ denote the conditional distribution of

$$\frac{\sqrt{n}\mu}{\sigma}\frac{1}{N}\sum_{j=1}^{n} c_{j,n}Y_j = (\mu/\bar{Y})U_n, \tag{3.12}$$

given $(T_n = m)$. We now state the main theorem.

Theorem 3.1. *Let $\sum_{j=1}^{n} a_{j,n} = 0$, $E(Y_1^6) < \infty$ and for some $M > 0$*

$$\sum_{j=1}^{n} \|a_{j,n}\|^3 < Mn. \tag{3.13}$$

Suppose for any $0 < K < L < \infty$, there exists a $\gamma = \gamma(K,L) < 1$ such that

$$\limsup_{n\to\infty} \sup_{K \le \|t\| < L} |d_n(t)| < \gamma. \tag{3.14}$$

If $m-np$ is bounded and if h is a real valued measurable function on $\boldsymbol{R}^k$ with $M_h < \infty$, then

$$\left|\int_{\boldsymbol{R}^k} h(x)(F_n^0(dx|m) - \tilde{\Psi}_n(x)dx)\right|$$
$$= o(M_h n^{-1/2}) + 0(w(h,\delta_n)), \tag{3.15}$$

for some $\delta_n = o(n^{-1/2})$.

Proof of Theorem 3.1 is differed to the Appendix.

In order to apply Theorem 3.1 to the sequential resampling procedures, let $X_1, X_2, \ldots$ be IID random vectors in $\boldsymbol{R}^k$ with mean vector η and dispersion Σ. Let H be a three times continuously differentiable function in a neighborhood of η and let $l(x)$ denote the vector of first order partial derivatives (gradient) of H. Suppose that the distribution of X_1 is strongly non-lattice, and $E\|X_1\|^3 < \infty$. Let $l(\eta) \neq 0$, $\theta^2 = l'(\eta)\Sigma l(\eta)$ a nd $\theta_n^2 = l'(\bar{X})\Sigma_n l(\bar{X})$, where Σ_n is the sample dispersion. By taking $a_{i,n} = (X_i - \bar{X})$, we can apply Theorem 3.1 to arrive at

Corollary 3.1. If $\gamma_3 = 1,\ \mu = \sigma,\ E(Y_1^6) < \infty$ and if $m - np$ is bounded, then

$$\sup_a \sqrt{n}|P(\sqrt{n}\Big(H\Big(\frac{1}{N}\sum_{i=1}^n X_i Y_i\Big) - H(\bar{X})\Big) \leq a\theta_n | T_n = m,\ X_1, \ldots, X_n)$$
$$-P(\sqrt{n}(H(\bar{X}) - H(\eta)) \leq a\theta) \to 0 \tag{3.16}$$

as $n \to \infty$, for almost all sample sequences $\{X_j\}$.

Corollary 3.2. Suppose the function H on $\boldsymbol{R}^k$ is three times continuously differentiable in a neighborhood of the origin and $H(0) = 0$. If $\gamma_3 = 1,\ E(Y_1^6) < \infty$, and $m - np$ is bounded. Then

$$\sup_a \sqrt{n}|P(\sqrt{n}H\Big(\frac{\mu}{\sigma N}\sum_{i=1}^n (X_i - \bar{X})Y_i\Big) \leq a\sqrt{l'(0)\Sigma_n l(0)} | T_n = m,\ X_1, \ldots, X_n)$$
$$-P(\sqrt{n}H(\bar{X} - \eta) \leq a\sqrt{l'(0)\Sigma l(0)})| \to 0 \tag{3.17}$$

as $n \to \infty$, for almost all sample sequences $\{X_j\}$.

Corollary 3.2 does not assume any relation between μ and σ, so it is applicable for a wide range of distributions for Y_1. In particular if Y_1 is negative binomial with parameters $r \geq 5$, and $P(Y_1 = 0) = (2 - (r/2) + (1/2)\sqrt{r(r-4)})^r$, then $\gamma_3 = 1,\ \mu = -r + 2r(4 - r + \sqrt{r^2 - 4r})^{-1}$ and $\sigma^2 = r/(r-4)$.

Proof of Corollary 3.2 is omitted as it is similar to the proof of Corollary 3.1.

Proof of Corollary 3.1. By expanding in Taylor series, we have

$$\sqrt{n}\Big(H(\frac{1}{N}\sum_{1}^{n} X_iY_i) - H(\bar{X})\Big)$$
$$= \frac{\mu}{\bar{Y}}U_n'V_nl(\bar{X}) + \frac{\mu^2}{\sqrt{n}\bar{Y}^2}U_n'V_nL_nV_nU_n + 0((\log n)^3 n^{-1}) \tag{3.18}$$

on $|U_n| < \log n$ and $|\bar{Y} - \mu| < \frac{\mu}{2}$.

Since the distribution of X_1 is strongly non-lattice, and $E\|X_1\|^3$ is assumed to be finite, conditions (3.13) and (3.14) hold for almost all sample sequences $\{X_i\}$. By Theorem 3.1 and by Lemma 3 of Babu and Singh (1984), we have

$$P(\sqrt{n}(H\left(\frac{1}{N}\sum_{i=1}^{n} X_iY_i\right) - H(\bar{X})) \le y\theta_n | T_n = m,\ X_1, \dots, X_n)$$
$$= \int_{-\infty}^{y} (1 + \frac{1}{\sqrt{n}}\gamma_3 q_0(x))\phi_1(x)\,dx + o(n^{-1/2}) \tag{3.19}$$

uniformly in y for almost all sample sequences, where q_0 is a polynomial. Similarly from the proofs of Theorems 20.8 and 24.2 of Bhattacharya and Ranga Rao (1986),

$$P(\sqrt{n}(H(\bar{X}) - H(\mu)) \le a\theta) = \int_{\infty}^{a} (1 + \frac{1}{\sqrt{n}} q_0(x))\phi_1(x)\,dx + o(n^{-1/2}) \tag{3.20}$$

uniformly in a. The result now follows from (3.19) and (3.20) as $\gamma_3 = 1$.

The most commonly used statistics, especially the studentized versions, are of the type

$$t_n = \sqrt{n}(H(\bar{X}) - H(\eta))/\nu\left(\frac{1}{n}\sum_{i=1}^{n}\lambda(X_i)\right), \tag{3.21}$$

where λ is a function on $\boldsymbol{R}^k \to \boldsymbol{R}^r$ and ν is a smooth real-valued function on $\boldsymbol{R}^r$. The classical Student's t is an example of this type of statistic. If X_i are univariate, then

$$t_n = \frac{\sqrt{n}(\bar{X} - \eta)}{s_n},$$

satisfies (3.21) with $H(x) = x$, $\lambda(x) = (x^2, x)$, $\nu(x, y) = \max(0, (x - y^2))^{1/2}$ and $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. The version corresponding to (3.21) under the Poisson scheme is generally of the type,

$$t_n(Y) = \sqrt{n}\left(H\left(\frac{1}{N}\sum_{i=1}^{n} X_iY_i\right) - H(\bar{X})\right)/\nu\left(\frac{1}{N}\sum_{i=1}^{n}\lambda(X_i)Y_i\right). \tag{3.22}$$

Corollary 3.3. Suppose Y_i is as in Corollary 3.1. Let

$$\nu(E(\lambda(X_1))) = \sqrt{l'(\eta)\Sigma l(\eta)}, \tag{3.23}$$

$$\nu(\bar{X}) = \sqrt{l'(\bar{X})\Sigma_n l(\bar{X})}, \tag{3.24}$$

and let $L(X_i)$ be a linearly independent sub collection of $(X_i, \lambda(X_i))$ with the property that all the coordinates of $(X_i, \lambda(X_i))$ can be expressed as linear combinations of those of $L(X_i)$. If the distribution of $L(X_1)$ is strongly non-lattice, $E\|L(X_1)\|^3 < \infty$, and if $m - np$ is bounded then

$$\sup_a \sqrt{n}|P(t_n(Y) \le a | T_n = m,\ X_1, \ldots, X_n) - P(t_n \le a)| \to 0, \tag{3.25}$$

as $n \to \infty$, for almost all sample sequences $\{X_j\}$.

Remark 1. If Y_1 is a Poisson random variable with mean 1, then $\mu = \sigma = \gamma_3 = 1$. If $m = [n(1 - e^{-1})] + 1$, then $0 \le m - np \le 1$. So Corollaries 3.1 and 3.3 are applicable for the basic Poisson scheme described in Section 2.

4. Appendix

To establish Theorem 3.1, we need some preliminary results. Let

$$F_n(x, r, m) = P(U_n \le x,\ N = r,\ T_n = m), \tag{4.1}$$

where $\nu \le x$ mean, the inequalities hold coordinate wise. Let

$$Z = (Y_1 - \mu,\ I_{\{Y_1 > 0\}} - p)',$$

and

$$\Psi_n(x, y) = (1 + \frac{1}{\sqrt{n}} Q_n(x, y))\phi_k(x)\varphi_0(y), \tag{4.2}$$

where φ_0 denotes the density of the bivariate normal distribution with zero mean vector and dispersion

$$\Sigma_0 = \begin{pmatrix} \sigma^2 & \mu q \\ \mu q & pq \end{pmatrix}. \tag{4.3}$$

Suppose

$$
\begin{aligned}
Q_n(x,y) &= \gamma_3 p_n(x) + \frac{1}{6}E\Big(Z'\Sigma_0^{-1}y\Big)^3 \\
&\quad - \frac{1}{2}(E((Z'\Sigma_0^{-1}y)(Y_1-\mu)^2pq + \sigma^2(q-p)I_{\{Y_1>0\}}))/\det\Sigma_0) \\
&\quad + \frac{1}{2}(\|x\|^2 - k)E\bigg(\Big(\frac{Y_1-\mu}{\sigma}\Big)^2 Z'\Sigma_0^{-1}y\bigg). \qquad (4.4)
\end{aligned}
$$

Since the support of Y_1 is assumed to include at least two non-zero values, it follows that Σ_0 is positive definite.

Proposition 4.1. *Under the conditions of Theorem 3.1,*

$$
\begin{aligned}
&\int_{R^k} h(x)(nF_n(dx,r,m) - \Psi_n(x,y_r,\omega_m)dx) \\
&= (o(M_h n^{-1/2}) + 0(\omega(h,\delta_n)))\varphi_0(y_n,\omega_m) \\
&+0(M_h n^{-1/2})\frac{1}{n}\sum_{i=1}^{n}\|a_{i,n}\|^3 E(Y_1^3 I_{\{Y_1\|a_{i,n}\|>\sqrt{n}\}}), \qquad (4.5)
\end{aligned}
$$

uniformly in

$$y_r = (r-n\mu)n^{-1/2},\ \omega_m = (m-np)n^{-1/2}, \qquad (4.6)$$

for some $\delta_n = o(n^{-1/2})$.

Remark 2. Under the conditions of Theorem 3.1,

$$\sup_{1\le j\le n}\|a_{j,n}\| = 0(n^{1/3}), \qquad (4.7)$$

so $\sup_{1\le j\le n}\|a_{j,n}\|n^{-1/2} \to 0$ as $n\to\infty$. Consequently

$$
\begin{aligned}
&\sup_{1\le j\le n} E(Y_1^3 I_{\{Y_1\|a_{j,n}\|>\sqrt{n}\}}) \\
&\quad = \sup_{1\le j\le n}\|a_{j,n}\|^3 n^{-3/2} E(Y_1^6 I_{\{Y_1\|a_{i,n}\|>\sqrt{n}\}}) \\
&\quad = o(n^{-1/2}).
\end{aligned}
$$

Hence the last term in (4.5) can be replaced by $o(M_h n^{-1})$.

The proof of Proposition 4.1 is similar to that of Theorem 1 of Babu and Bai (1996). It uses truncation arguments in the proof of Theorem 20.8 of Bhattacharya and Ranga Rao (1986). The Proposition follows from Lemmas 4.1, 4.2 and 4.3 that

are stated below, and Lemma 4 of Babu and Bai (1996). Lemmas 4.1, 4.2 and 4.3 are modified versions of the first three lemmas of Babu and Bai (1996). The measure J in Babu and Bai (1996) is assumed to satisfy $\int \|x\|^{k+14} dJ(x) < \infty$.

Before stating the lemmas, we introduce some notation. For non-singular integral vector $\alpha' = (\alpha_1, \dots, \alpha_s)$ and $z' = (z_1, \dots, z_s) \in \boldsymbol{R}^s$, we write

$$|\alpha| = \alpha_1 + \ldots + \alpha_s, \; z^\alpha = z_1^{\alpha_1} \ldots z_s^{\alpha_s} \text{ and } D^\alpha = D_1^{\alpha_1} \ldots D_s^{\alpha_s},$$

where D_j denotes the partial derivative with respect to the j-th coordinate.

Lemma 4.1. *Let g be a real valued function on $\boldsymbol{R}^k \times \boldsymbol{Z}^j$ satisfying*

$$\sum_{\underset{\sim}{m} \in \boldsymbol{Z}^j} \int_{\boldsymbol{R}^k} (1 + \|x\|)^{s+k+1} |g(x, \underset{\sim}{m})| dx < \infty,$$

for some non-negative s. Then there exists a constant $c(k)$ depending only on k such that, for all $\underset{\sim}{m} \in \boldsymbol{Z}^j$,

$$\int_{\boldsymbol{R}^k} (1 + \|x\|^s) |g(x, \underset{\sim}{m})| dx$$
$$\leq c(k) \max_{|\alpha| \leq 1+k+s} \int_G \left(\int_{\boldsymbol{R}^k} |D^\alpha \hat{g}(t, \nu) \right) |dt) d\nu,$$

where $G = [-\pi, \pi]^j$ and $\hat{g}$ denotes the Fourier transform of g.

Proof of Lemma 4.1 is similar to that of Lemma 1 of Babu and Bai (1996) and hence omitted.

Lemma 4.2. *Let Y^0 and ε^0 have the same distributions as that of Y and ε respectively. Suppose (Y, ε) and (Y^0, ε^0) are independent. Then*

$$\frac{1}{n} \sum_{j=1}^{n} |E(e^{it' a_{j,n} Y + i\nu' \varepsilon})|^2 \leq E|d_n(t(Y - Y^0))| \tag{4.8}$$

Proof of Lemma 2. The left side of (4.8) is same as

$$\frac{1}{n} \sum_{j=1}^{n} E(e^{it' a_{j,n}(Y - Y^0)} e^{i\nu'(\varepsilon - \varepsilon^0)})$$
$$= E(d_n(t(Y - Y^0)) e^{i\nu'(\varepsilon - \varepsilon^0)})$$
$$\leq E|d_n(t(Y - Y^0))|.$$

Lemma 4.3. *Let $E(Y_1^3) \leq M_1$ and the smallest eigenvalue of Σ_0 is bounded below by $\sigma_1 > 0$. Suppose (3.13) holds. Then for $\|t\| \leq n^{-1/2}\log n$ and $Mn^{-1/2}\log n < \|\nu\| \leq \pi$, we have for any $C \subset \{1, \ldots, n\}$,*

$$\left|\prod_{j\in C} E(e^{it'a_{j,n}Y_1 + i\nu' Z})\right| \leq k_1 \exp(b - k_2(\log n)^2)$$

where b denotes the number of integers in C and $k_1, k_2 > 0$ are positive constants depending only on M, M_1, b and σ_1.

Proof of Lemma 3. Since

$$|e^{-i\omega\mu - i\nu p}E(e^{i\omega Y_1 + iuI_{\{Y_1>0\}}}) - 1 + \frac{1}{2}(\omega, u)\Sigma_0(\omega, u)'| \leq \frac{1}{6}\|(\omega, u)\|^3 M_1,$$

there exists a $0 < \delta < \pi/8$ and $\Delta_1 > 0$ depending only on σ_1 and M_1 such that

$$\frac{1}{2} \leq |E(e^{i(\omega,u)Z})| \leq 1 - \Delta_1(\omega^2 + u^2),$$

whenever $|\omega| < 4\delta$ and $|u| < 4\delta$. Suppose $\|t\| \leq n^{-1/2}\log n$, and $Mn^{-1/2}\log n < \|\nu\| < \delta$. Then for some $\Delta > 0$

$$\begin{aligned}
\left|\prod_{j=1}^{n} E(e^{it'a_{i,n}Y_1 + i\nu' Z})\right| &\leq \prod_{j=1}^{n}(1 - \Delta((t'a_{j,n})^2 + \|\nu\|^2) \\
&\leq e^{-\Delta n\|t'V_n\|^2 - n\Delta\|\nu\|^2} \\
&\leq e^{-\Delta M(\log n)^2}.
\end{aligned} \tag{4.9}$$

Further observe that

$$\begin{aligned}
|E(e^{i\omega Y_1 + i\nu I_{\{Y_1>0\}}})| &= |q + pe^{i\nu}E(e^{i\omega Y_1}|Y_1 > 0)| \\
&\leq q + p|E(e^{i\omega Y_1}|Y_1 > 0)| \\
&\leq 1 - \gamma_0 < 1,
\end{aligned}$$

for some $\gamma_0 > 0$, whenever $\frac{\delta}{2} < |\omega| < \pi + \delta$. Hence

$$|E(e^{it'a_{j,n}Y_1 + i\nu' z})| \leq 1 - \gamma_0,$$

whenever $\|t\| \leq n^{-1/2}\log n$ and $\delta \leq \|\nu\| \leq \pi$. This completes the proof of Lemma 4.3.

Proof of Theorem 3.1. Let H_n denote the indicator function of $(2|\frac{1}{n}\sum_{j=1}^{n} Y_i - \mu| > \mu)$. Suppose $m - np$ is bounded. By Markov inequality, $E(H_n) = 0(n^{-3})$. Hence

$$\begin{aligned} E(H_n|T_n = m) &= E(H_n)/P(T_n = m) \\ &= 0(n^{-3})(P(T_n = m))^{-1} \\ &= 0(n^{-5/2}). \end{aligned}$$

Clearly

$$\begin{aligned} E(Y_1 H_n N^{-1}|T_n = m) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i H_n N^{-1}|T_n = m) \\ &= \frac{1}{n}E(H_n|T_n = m) \\ &= 0(n^{-7/2}). \end{aligned}$$

Consequently, there exists a constant $M_2 > 0$, such that

$$\begin{aligned} &M_2 E(h(\sqrt{n}\mu\sum_{j=1}^{n} c_{j,n}Y_j/(N\sigma))|T_n = m) \\ &\le E(H_n|T_n = m) + n^{3/2}E(\|\sum_{j=1}^{n} c_{j,n}Y_j N^{-1}H_n\|^3|T_n = m) \\ &= 0(n^{-7/2}) + n^{3/2}E(\sum_{j=1}^{n} \|c_{j,n}\|^3 Y_j N^{-1} H_n|T_n = m) \\ &= 0(n^{-7/2}) + 0(n^{3/2}E(H_n|T_n = m))| \\ &= 0(n^{-1}). \end{aligned} \tag{4.10}$$

By (3.12), (4.1) and (4.6), we have

$$F_n^0(x|m) = \sum_{j=1}^{\infty} F_n(x(1 + (y_j/\mu\sqrt{n})),\ y_j, \omega_m)/P(T_n = m),$$

and by Proposition 4.1,

$$\begin{aligned} &|\frac{1}{\sqrt{n}}\sum_{j:2|y_j|<\mu\sqrt{n}} h(x/(1 - (y_j/\mu\sqrt{n})) \\ &\quad\times (nF_n(dx, y_j, \omega_m) - \Psi_n(x, y_j, \omega_m)dx| \\ &= (o(M_h n^{-1/2}) + 0(\omega(h, \delta_n)))\frac{1}{\sqrt{n}}\sum_{|y_j|\le\mu\sqrt{n}} \varphi_0(y_j, \omega_m) + o(M_h n^{-1/2}) \\ &= o(M_h n^{-1/2}) + 0(\omega(h, \delta_n)). \end{aligned} \tag{4.11}$$

By Theorem 13 on local Edgeworth expansions on pages 205-206 of Petrov (1975), we have

$$\sqrt{pqn}P(T_n = m) = \phi_1(\omega_m/\sqrt{pq})\left(1 + \frac{q-p}{6\sqrt{npq}}\left(\left(\frac{\omega_m}{\sqrt{pq}}\right)^3 - \frac{\omega_m}{\sqrt{pq}}\right)\right) + o(n^{-1/2}). \tag{4.12}$$

The Theorem now follows from (4.11) and (4.12), as in the proof of Theorem 6 of Babu and Bai (1996).

REFERENCES

[1] Babu, G. J. and Bai, Z. D. (1996). Mixtures of global and local Edgeworth expansions and their applications. *J. Multivariate Analysis*, **59**, 282-307.

[2] Babu, G. J. and Singh, K. (1983). Inference on means using the bootstrap. *Annals of Statistics*, **11**, 999-1003.

[3] Babu, G. J. and Singh, K. (1984). On the term Edgeworth correction by Efron's bootstrap. *Sankhyā*, **46A**, 219-232.

[4] Babu, G. J. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from finite populations. *Journal of Multivariate Analysis*, **17**, 261-278.

[5] Babu, G. J. and Singh, K. (1989). On Edgeworth expansions in the mixture cases. *Annals of Statistics*, **17**, 443-447.

[6] Bai, Z. D. and Rao, C. R. (1992). A note on Edgeworth expansion for ratio of sample means, *Sankhyā*, **54A**, 309-322.

[7] Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Annals of Statistics*, **6**, 434-451.

[8] Bhattacharya, R. N. and Ranga Rao, R. (1986). *Normal Approximations and Asymptotic Expansions*, Krieger Publishing Company, Florida.

[9] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Ann. Statist.*, **7**, 1-26.

[10] Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures, *Ann. Prob.*, **18**, 851-869.

[11] Hall, P. and Mammen, E. (1994). On general resampling algorithms and their performance in distribution estimation, *Ann. Statist.*, **24**, 2011-2030.

[12] Pathak, P. K. (1962). On simple random sampling with replacement, *Sankhyā*, **24A**, 283-302.

[13] Petrov, V. V. (1975). *Sums of Independent Random Variables*, English translation, Springer-Verlag, New York.

[14] Rao, C. R., Pathak, P. K. and Koltchinskii, V. I. (1997). Bootstrap by sequential resampling. *Journal of Statistical Planning and Inference*, **64**, 257-281.

[15] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187-1195.

GUTTI JOGESH BABU, DEPARTMENT OF STATISTICS, 326 THOMAS BUILDING, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PA 16902.

P. K. PATHAK, MICHIGAN STATE UNIVERSITY AND UNIVERSITY OF NEW MEXICO.

C. R. RAO, DEPARTMENT OF STATISTICS, 326 THOMAS BUILDING, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PA 16902.